DOCUMENT RESUME

ED 441 861 TM 030 897

AUTHOR Roussos, Louis; Nandakumar, Ratna; Cwikla, Julie
TITLE DIF Assessment of CAT Data: Kernel-Smoothed CATSIB.
PUB DATE 2000-04-00
NOTE 20p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; *Item Bias; Item Response Theory; Nonparametric Statistics; *Sample Size; Simulation
IDENTIFIERS *CATSIB Computer Program; Item Bias Detection; *Kernel Method; Smoothing Methods

ABSTRACT

CATSIB is a differential item functioning (DIF) assessment methodology for computerized adaptive test (CAT) data. Kernel smoothing (KS) is a technique for nonparametric estimation of item response functions. In this study an attempt has been made to develop a more efficient DIF procedure for CAT data, KS-CATSIB, by combining CATSIB with kernel smoothing. A correction for smoothing in boundaries is also implemented. It is hoped that such a methodology could provide a more powerful DIF technique for smaller samples while enhancing the interpretation of local DIF analyses. A simulation study was conducted to investigate the DIF estimation bias of KS-CATSIB in comparison to CATSIB with small samples. Sixteen DIF items varying in difficulty and discrimination were considered for this purpose. A sample of 500 examinees was used in the reference group and a sample of 250 examinees was used in the focal group. Preliminary results show that the correction for smoothing in boundaries, even though effective in reducing the bias in estimation, is still larger for KS-CATSIB in comparison with CATSIB. Therefore, DIF estimates associated with KS-CATSIB are statistically biased and would lead to high Type I error rates. Further modifications of KS-CATSIB are necessary before the program is ready for full implementation. (Contains 3 tables and 13 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 861

DIF Assessment of CAT Data: Kernel-Smoothed CATSIB

Louis Roussos
Law School Admission Council

Ratna Nandakumar and Julie Cwikla
University of Delaware

BEST COPY AVAILABLE

TM030897

Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 2000.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

R Nandakumar

1

DIF Assessment of CAT Data: Kernel-Smoothed CATSIB
ABSTRACT

CATSIB is a DIF assessment methodology for Computerized adaptive test data. Kernel smoothing is a technique for nonparametric estimation of item response functions. In this study an attempt has been made to develop a more efficient DIF procedure for CAT data, KS-CATSIB, by combining CATSIB with kernel smoothing. A correction for smoothing in boundaries is also implemented. It is hoped that such a methodology could provide a more powerful DIF technique for small samples while enhancing the interpretation of local DIF analyses.

A simulation study was conducted to investigate the DIF estimation bias of KS-CATSIB in comparison to CATSIB with small samples. Sixteen DIF items varying in difficulty and discrimination were considered for this purpose. A sample of 500 examinees was used in the reference group and a sample of 250 examinees was used in the focal group. Preliminary results showed that the correction for smoothing in boundaries, even though effective in reducing the bias in estimation, the bias in estimation is still larger for KS-CATSIB in comparison to CATSIB. Therefore, DIF estimates ($\overline{\hat{\beta}}$s) associated with KS-CATSIB are statistically biased and would lead to high Type I error rates. Further modifications of KS-CATSIB are necessary before the program is ready for full implementation.

DIF Assessment of CAT Data: Kernel-Smoothed CATSIB

CATSIB (Nandakumar & Roussos, 1997; Roussos, 1996) is a DIF detection procedure for assessing DIF of computerized adaptive test (CAT) items. It can be used for DIF assessment at the pretest stage and continue with DIF monitoring at the operational stage, using the combined pretest and operational data, if necessary. Kernel smoothing is a technique for nonparametric estimation of an IRF (Ramsay, 1991). It is a computationally simple technique for estimating IRFs and does not impose any assumptions on the data. However, merely estimating reference (R) and focal (F) group IRFs does not result in a DIF hypothesis test. Hence by combining the two procedures it is hoped that a more efficient DIF detection procedure can be obtained. Douglas, Stout, and DiBello (1996) have combined the kernel-smoothing with SIBTEST to obtain a new DIF statistic for detecting DIF of paper-and-pencil tests. They have shown that kernel-smoothed SIBTEST has the advantage of efficient detection of local DIF and increased power for the detection of differentially functioning items.

Pretest sample sizes for computerized adaptive tests, for a variety of reasons, tend to be smaller than those for paper-and-pencil pretest items. Thus, it is important to develop DIF detection procedures that have maximum detection power even for small samples. Therefore, the purpose of this study is to develop an effective DIF detection procedure for CAT data by combining two statistical methodologies: CATSIB and kernel-smoothing item response function estimation. The resulting procedure will be referred to as KS-CATSIB. In this paper the DIF estimation bias of KS-CATSIB in comparison to CATSIB is studied for small sample sizes.

In what follows, CATSIB procedure will be first described followed by the Kernel-smoothed IRFs, followed by KS-CATSIB. Following this, the simulation design and results will be described.

CATSIB

Let θ denote the proficiency of an examinee on the construct measured by the test. An item is said to display DIF if the reference group test takers and the focal group test takers, matched on proficiency level θ, do not have the same probability of correct response on the item. Let $DIF(\theta)$ be defined as the magnitude of DIF in a studied item at proficiency level θ. A common way to model DIF as given in Shealy and Stout (1993) is:

$$DIF(\theta) = P_R(\theta) - P_F(\theta), \tag{1}$$

where $P_R(\theta)$ and $P_F(\theta)$ denote item response functions of the studied item for the R and F groups respectively at proficiency level θ. DIF, denoted by β, is defined as the average of $DIF(\theta)$ over θ given as:

$$\beta = \int DIF(\theta) f(\theta) d\theta, \tag{2}$$

where $f(\theta)$ is an appropriate density function on θ such as that for the combined R and F groups. Hence, the null hypothesis for DIF hypothesis testing is stated as

$$H_0 : \beta = 0.$$

An important aspect of any DIF procedure is to select R and F test takers that are matched on ability before comparing their performance on the studied item. An obvious choice for this purpose is observed test scores or ability estimates. However, it is a well known fact that some type of correction is necessary to observed estimates of ability to avoid type I errors. This is because, it is often observed that there is a stochastic ordering of ability distributions on the intended ability of R and F groups, with R group mean higher than the F group mean. Therefore matching on the observed (estimated) score could result in falsely detecting DIF in favor of the group with higher mean ability, thus inflating the Type I error. Hence a correction is commonly adopted to correct for this statistical bias.

CATSIB, following the tradition of SIBTEST (Shealy & Stout, 1993), employs a "regression correction" to correct for ability differences in R and F groups. Instead of matching examinees on estimated θ (denoted by $\hat{\theta}$), CATSIB matches examinees in R and F groups on an estimate of expected true values of θ, for group g (R or F) as given by (see Nandakumar and Roussos (in press) for details)

$$\hat{\theta}^* = E_g[\theta|\hat{\theta}]. \tag{3}$$

In order to compute an estimate of DIF (denoted by $\hat{\beta}$), the observed range of real-valued variable $\hat{\theta}^*$ is divided into n equal intervals. Examinees are then classified into one of the n intervals based on their values of $\hat{\theta}^*$. An estimate of DIF is then given by

$$\hat{\beta} = \sum_{k=1}^{n} [\hat{P}_{R,k} - \hat{P}_{F,k}] \hat{p}_k, \tag{4}$$

where $\hat{P}_{g,k}$ is the observed proportion of group g test takers in ability interval k who got the studied item right, and $\hat{p}_k$ is the observed proportion of R and F test takers who were classified into interval k.

Because the number of intervals was arbitrary, an initial approach we took was to have the computer program automatically determine the number of intervals. To ensure stable statistical estimation, an interval was required to have a minimum of three test takers from each of R and F for that interval to be included in the calculation of $\hat{\beta}$. All intervals with fewer than this minimum number were not used. Thus, it was important to carefully choose the number of intervals. If too many intervals were to be used, the intervals could become so sparsely populated with test takers that too many intervals (and, thus, too many test takers) could be eliminated from the statistic calculation resulting in a powerless statistic. On the other hand, if too few intervals were to be used, the test statistic could become overly sensitive to impact and its Type 1 error could become unacceptably inflated. (In the extreme case of a single interval, the statistic would reduce to being purely a measure of impact.) To strike a balance between these two extremes, CATSIB was programmed to automatically start with an arbitrarily large number of ability intervals (80) and to then monitor how many test takers would be eliminated due to the throwing out of sparse cells. If more than 7.5% of either the R or F test takers would be eliminated, CATSIB automatically decreases the number of cells until the number of test takers eliminated from each group becomes less than or equal to 7.5%. However, the minimum number of ability intervals was set at 20, even if this meant that the number of test takers eliminated from one or both of the groups sometimes exceeded 7.5%.

Subsequently, a systematic investigation into the number of intervals to be used to obtain optimal estimates of DIF and hypothesis testing was conducted (Roussos, Nandakumar, & Cwikla, 1999). Two scaling methods were used: normal and percentile, to classify examinees into intervals based on $\hat{\theta}^*$. The results revealed that DIF can be estimated accurately with as few as 10 intervals with minimal bias in either the estimation of DIF or the estimated standard error of the statistic. Percentile scale in general provided accurate results with fewer number of intervals than the normal scaling.

The test statistic to test the null hypothesis of no DIF is given by

$$B = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})}. \tag{5}$$

The standard error for $\hat{\beta}$ is estimated based on the observed variance of the studied item responses in each ability interval:

$$\hat{\sigma}(\hat{\beta}) = \sqrt{\sum_{k=1}^{n} \left[\frac{\hat{\sigma}^2_{R,k}(Y)}{n_{R,k}} + \frac{\hat{\sigma}^2_{F,k}(Y)}{n_{F,k}} \right] \hat{p}^2_k}. \tag{6}$$

where Y denotes the response to the studied item, $\hat{\sigma}^2_{g,k}(Y)$ is the observed variance of Y in ability interval k for group g, and $n_{g,k}$ is the number of g group test takers in interval k. The null hypothesis of no DIF is rejected at level α if the statistic B exceeds the $100(1-\alpha)$ percentile obtained from the standard normal table.

The estimate $\hat{\beta}$ serves as an index of the amount of DIF present in the item. For example, it is possible that an item may exhibit statistically significant DIF but the degree of DIF may not be practically meaningful in terms of how it affects the performance of test takers in the two groups. Thus, $\hat{\beta}$ can be very useful in assessing the degree of DIF practically. It can be seen from Equation 4 that $\hat{\beta}$ estimates the average difference between R and F in percent chance of a correct response (conditional on $\hat{\theta}^*$) on the studied item.

Kernel – Smoothed Item Response Functions (7)

This section provides a brief overview of kernel-smoothing estimation of unknown regression functions. Kernel-smoothing for estimation of item response functions was first introduced by Ramsay (1991) and subsequently used for assessing DIF of paper-and-pencil tests by Douglas, Stout, and DiBello (1996). It is a general technique for estimating unknown nonparametric regression functions of $E[Y|X=x]$. In nonparametric estimation, no assumptions are made regarding the shape of the unknown regression function such as linear or nonlinear.

Nadaraya (1964) and Watson (1964) proposed a method to estimate a nonparametric regression function $E[Y|X=x]$ as

$$f_n(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{b}) Y_i}{\sum_{i=1}^{n} K(\frac{x-x_i}{b})} \tag{8}$$

Here n is the sample size and K is a weight function that is symmetric, nonnegative and approaches zero as its argument becomes further from zero. The function K has support on the interval [-1,1]. The user-specified constant b is a bandwidth parameter which controls the amount of smoothing. It can be seen that $f_n(x)$ is a smoothly weighted average of the Y_i, where the weights are determined by the kernel function $K(x)$ and the bandwidth b. Also, the further x_i is from x, the less weight is given to the corresponding value Y_i in the estimation of $f_n(x)$.

It has been shown that the accuracy of estimation of $f_n(x)$ is very sensitive to the bandwidth b. If the bandwidth is very small, only points with ordinates very close to

the x value where $f_n(x)$ is estimated are given much weight. Although, this weighted averaging within a narrow window helps in controlling the local behavior of the regression function, it excludes data and as a result the estimate of the regression curve is not smooth. A large bandwidth, on the other hand, includes more data, thereby compromising the local behavior with a smoother regression curve. A way to compromise is to choose a bandwidth that is a function of sample size. A rule of thumb for the bandwidth is to set $b = Cn^{-1/5}$, where C is chosen on the basis of past experience. Douglas, Stout, and DiBello (1996) in their study chose $C = 0.7$, and $K(x) = 1 - x^2$ for $-1 \leq x \leq 1$.

Kernel-Smoothed CATSIB

KS-CATSIB is obtained by incorporating kernel-smooth estimates of IRFs in CATSIB instead of using mutually exclusive cells corresponding to discrete ability intervals. In this way, a more natural use of the matching criterion is utilized and avoids the problem of sparse cells, minimum number of examinees per cell, etc. The regression correction implemented in CATSIB is unaffected by the use of kernel smoothing and thereby can still be used to control for the type I error rates. It is hoped that the use of kernel-smoothed IRFs will enhance the assessment of local DIF and increas the power with small samples.

Steps in implementing KS-CATSIB are described below. It is assumed here that the items used for obtaining the ability estimates (in order to match examinees from the reference and focal groups) are disjoint from the studied item(s).

Step 1. For each examinee obtain an estimate of ability ($\hat{\theta}_i$) by adaptively administering pre calibrated items.

Step 2. Perform the regression correction on ability estimates (denoted by $\hat{\theta}_i^*$) for the focal and reference group examinees separately.

Step 3. Divide the ability range (taken over both groups combined) into N equally spaced design points.

Step 4. Compute the kernel-smoothed estimate of item response function for the studied item for group g at each of the design points according to

$$\hat{P}_g(\theta_k) = \frac{\sum_{i=1}^{n_g} K(\frac{\theta - \hat{\theta}_{ig}^*}{h(n_g)}) Y_{ig}}{\sum_{i=1}^{n_g} K(\frac{\theta - \hat{\theta}_{ig}^*}{h(n_g)})}, \quad k = 1, \ldots, N$$

where
$K(x) = 1 - x^2$, $-1 \leq x \leq 1$
$h(n_g) = 0.7n^{-1/5}$ is the smoothing parameter or the band width,
n_g is the sample size in group g, and
Y_{ig} is the response to the studied item by the ith examinee of group g.

Step 5. Estimate local DIF $\beta(\theta_k)$ by

$$\hat{\beta}(\theta_k) = \hat{P}_R(\theta_k) - \hat{P}_F(\theta_k)$$

Step 6. Estimate global DIF β by

$$\hat{\beta} = \hat{E}[\hat{P}_R(\theta) - \hat{P}_F(\theta)]$$

as estimated by

$$\hat{\beta} = \frac{\sum_1^{n_R+n_F}(\hat{P}_R(\hat{\theta}_i^*) - \hat{P}_F(\hat{\theta}_i^*))}{n_R + n_F}$$

Boundary modifications for kernel regression

Although the kernel window is symmetric in the middle of the ability range, it becomes asymmetric for evaluating points that are less than a bandwidth away from either boundary. In this case, the kernel-smoothing estimation procedure must be adjusted. Our initial corrections for boundary problem is made following Rice (1984) as follows.

- IF $\theta < h(n_g)$. That is, θ is in the left boundary (below the band width).

Let $\rho = \frac{\theta}{h(n_g)}$
The modified IRF is given by

$$\hat{P}_g'(\theta) = \hat{P}_g(\theta; h(n_g)) + \beta[\hat{P}_g(\theta; h(n_g)) - \hat{P}_g(\theta; \alpha h(n_g))],$$

where
$\beta(\rho) = \frac{R(\rho)}{\alpha R(\frac{\rho}{\alpha}) - R(\rho)}$, $\alpha = 2 - \rho$, and $R(u) = \frac{W_1(u)}{W_0(u)}$

$W_0(u) = \int_{-1}^{u} K(v)\, dv = u(1 - \frac{u^2}{3}) - \frac{2}{3}$
$W_1(u) = \int_{-1}^{u} vK(v)\, dv = \frac{u^2}{2}(1 - \frac{u^2}{2}) - \frac{1}{4}$

• IF $\theta > 1 - h(n_g)$. That is, θ is in the right boundary (above the band width).

Let $\rho = \frac{1-\theta}{h(n_g)}$
The modified IRF is given by

$$\hat{P}'_g(\theta) = \hat{P}_g(\theta; h(n_g)) + \beta[\hat{P}_g(\theta; h(n_g)) - \hat{P}_g(\theta; \alpha h(n_g))],$$

where
$\beta(\rho) = \frac{R(\rho)}{\alpha R(\frac{\rho}{\alpha}) - R(\rho)}$, $\alpha = 2 - \rho$, and $R(u) = \frac{W_1(u)}{W_0(u)}$

$W_0(u) = \int_u^1 K(v)\, dv = \frac{2}{3} - u(1 - \frac{u^2}{3})$
$W_1(u) = \int_u^1 vK(v)\, dv = \frac{1}{4} - \frac{u^2}{2}(1 - \frac{u^2}{2})$

The Simulation Study

To evaluate the performance of KS-CATSIB with CATSIB, a simulation study was designed. Small sample sizes, typically occuring in pretest situations were considered. Sample sizes of 500 and 250 were used for the reference group, while sample sizes of 500, 250, and 125 were used for the focal group.

Examinee abilities were generated from normal distributions. The R and F groups had the same standard deviation of 1. However the means of the R and F distributions differed by one standard deviation. That is, the impact level ($d_T = \mu_{\theta_R} - \mu_{\theta_F}$) was set to 1.0. Two different distributional situations were used with $d_T = 1$: $\mu_R = 0.5$ and $\mu_F = -0.5$ when R and F were of the same size; and $\mu_R = 0.67$ and $\mu_F = -0.33$ when R was twice the size of F.

In generating the item parameters of the item pool for the simulated operational items, the goal was to generate parameters that closely resembled those estimated from real data. To this end, the descriptive properties of the item parameter estimates from over 700 operational LSAT items were observed. It was found that the item discrimination parameters generally ranged from .4 to 1.1 for low difficulty level items and from .4 to 1.7 for high difficulty items, and that the distributions were positively skewed. The vast majority of the item difficulty parameters were observed to range between -3 and 3 and followed approximately the standard normal distribution. The simulated difficulty level parameters were, thus, generated from the standard normal distribution. The simulated discrimination parameters were generated from one of two lognormal distributions,

depending on the difficulty level of the item. The simulated lower asymptote parameters were independently generated from a uniform distribution with a range between .12 and .22 to approximate those from the actual LSAT data. The precise distributions used for the item parameters are described below.

$$log(a) \sim \text{normal } (-.357, .25) \text{ for } b \leq -1 \text{ with range } .4 \leq a \leq 1.1$$

$$log(a) \sim \text{normal } (-.223, .34) \text{ for } b > -1 \text{ with range } .4 \leq a \leq 1.7$$

$$b \sim N(0, 1) \text{ with range } -3 \leq b \leq 3$$

$$c \sim \text{U}(.12, .22)$$

DIF was introduced in the studied items through differences in the difficulty parameters between R and F using the following model for DIF:

$$\text{DIF} = \hat{\beta} = \int [P_R(\theta) - P_F(\theta)] f(\theta) d\theta, \tag{9}$$

where

$$P_g(\theta) = c + \frac{1 - c}{1 + exp[-1.7a(\theta - b_G)]}, \; g = R \text{ or } F \tag{10}$$

Sixteen different DIF items were considered. They varied in the amount of latent DIF (β) in each item and in the difficulty and discrimination parameters. Three DIF levels were considered: 0, .05 and .1. An item with β value of 0 indicates the null case of no DIF. An item with a β value of 0.050 indicates the lowest value of medium DIF, and an item with a β value of 0.100 indicates the lowest value of high DIF. There were 6 no DIF items ($\beta = 0$); 5 medium DIF items ($\beta = 0.050$); and 5 high DIF items ($\beta = 0.100$). Within each category of DIF, items varied in difficulty and discrimination: medium discrimination (.8) and medium difficulty (0); low discrimination (.4) and low difficulty (-1.5); low discrimination (.4) and high difficulty (1.5); moderate discrimination (1.0) and low difficulty (-1.5); and high discrimination (1.4) and high difficulty (1.5). One more extreme item was included for $\beta = 0$ case, high discrimination (1.4) and low difficulty (-1.5). This item was included only in the null DIF case because previous studies had shown that high discriminating and easy items have a tendency for impact-induced Type I error inflation (Roussos & Stout, 1996; Allen & Donoghue, 1996). These items are listed in Table 1. Items 1 to 6, in Table 1, are easy and low discriminating, items 7 to 9 are of average difficulty and discrimination, and items 10 to 15 have high difficulty and high discrimination. Item 16 is the rare item that is very easy and has high discrimination.

For each examinee, item responses to DIF items were computed using the three-parameter logistic function given by,

$$P_g(\theta) = c + \frac{1-c}{1+exp[-1.7a(\theta - b_g)]},\ g = R \text{ or } F \tag{11}$$

where a and b denote the discrimination index and difficulty index, respectively, of the DIF item under investigation. The guessing level c was set to .17, which is the average of all estimated c parameters of the LSAT item bank; and θ is the true (generated) ability level of the examinee.

Each simulated examinee was adaptively administered 25 operational items (a value typical of operational CATs), and linearly administered all 16 DIF items. An estimate of ability was obtained from the examinee responses to the operational items. The ability estimates of test takers were determined using a standard maximum-information CAT design described as follows.

The ability scale from -2.25 to 2.25 was divided into 37 equal intervals in increments of 0.125. For each item i, item information, $I_i(\theta)$, was computed at the θ values corresponding to the midpoints of the 37 intervals using the following formula (Hambleton, Swaminathan, & Rogers, 1991, p. 91):

$$I_i(\theta) = \frac{(1.7a_i)^2(1-c_i)}{[c_i + exp(1.7a_i(\theta - b_i))][1+exp(-1.7a_i(\theta - b_i))]^2}$$

where a_i, b_i, and c_i denote discrimination, difficulty, and lower asymptote parameters of item i respectively. At each θ level the pool of operational items was sorted according to the item information values from lowest to highest and saved in a separate table. This table was used during the simulations to select items with the highest information at a given θ-level.

To prevent items from becoming overexposed, an exposure control method was incorporated (Kingsbury & Zara, 1989). Accordingly, the first item to be administered to a simulated test taker was randomly selected from the 10 items with highest information values at $\theta = 0$ (the starting value for all simulated test takers). The second item was randomly selected from the 9 best items at the new estimate of θ. The third item was randomly selected from the 8 best items, and so on until, beginning with the 10th item, the item with the highest information was selected (unless, of course, the item had already been administered to that simulated test taker, in which case the next best item was selected).

After administering each item in this manner to each test taker, the simulated test taker's response (right/wrong) was determined, and the simulated test taker's estimated ability, $\hat{\theta}$, was updated using Owen's Bayesian sequential scoring (Owen, 1969). After all 25 items were administered, a Bayesian modal score was calculated and was used as the final ability estimate ($\hat{\theta}$).

After applying the CATSIB regression correction to the ability estimates (denoted by $\hat{\theta}_i^*$), kernel-smoothed IRFs were obtained for each DIF item and the DIF estimate ($\hat{\beta}$) was computed for CATSIB and KS-CATSIB. This process was repeated 100 to 400 times for each DIF item and the average DIF estimate ($\overline{\beta}$) was computed over 100 (or 400) trials. These results were compared with true DIF (β) values. The results for all 16 DIF items are tabulated in Tables 2 and 3.

RESULTS

Tables 2 and 3 summarize preliminary results with sample sizes of 500 for the reference group and 250 for the focal group.

Table 2 compares results of KS-CATSIB with and without the Rice correction for the boundaries. The body of the table provides average DIF estimates ($\overline{\beta}$) over 100 trials along with the standard errors in the parentheses. The general pattern of results show that there is a slight overestimation of DIF by both procedures, and $\overline{\beta}$s with the Rice correction are closer to the true values of DIF than $\overline{\beta}$s without the Rice correction (constant bandwidth). The average bias ($\beta - \overline{\beta}$) for KS-CATSIB with the Rice correction was -.0104, and the average bias without the Rice correction was -.0136. Similarly, the standard deviation in bias with Rice correction was lower (.0108) than without the Rice correction (.0120). Therefore, generally, the Rice correction for boundaries appears to be effective in reducing the statistical bias in kernel-smoothing estimation at the boundaries. In few instances where the bias was more with rice correction happens to be for high difficult or high difficult and high discrimination items. That is, for very difficult and/or discrimination items, adjusting the bandwidth at the boundaries is causing more harm. It appears that for such items, extreme sparseness of examinees at one of the boundaries is further inflating the bias with Rice correction. Further research is needed to examine the Rice correction and also to identify other boundary corrections that work for all types of items.

Table 3 compares DIF estimates ($\overline{\beta}$s) for KS-CATSIB and CATSIB for 100 trials

and 400 trials. As expected, DIF estimates ($\overline{\hat{\beta}}$) with 400 trials are slightly closure to true values of β, than those with 100 trials. In all cases, DIF estimates of CATSIB are closer to true values than those with KS-CATSIB. That is, the bias in DIF estimation ($\beta - \overline{\hat{\beta}}$) is higher for KS-CATSIB. This is further augmented for high difficult and discrimination items. For the case of 400 trials, the average bias over 16 items for CATSIB was $-.0045$, while the average bias for KS-CATSIB was $-.0164$. The standard errors of bias for CATSIB was also lower (.0108), than KS-CATSIB (.0120). Therefore, $\overline{\hat{\beta}}$s associated with KS-CATSIB are statistically biased and would lead to high Type I error rates. Further modifications of KS-CATSIB are necessary before the program is ready for full implementation.

Conclusion

In this study an attempt has been made to develop a more efficient DIF procedure for CAT data by combining CATSIB with kernel smoothing. It is hoped that such a methodology could provide a more poweful DIF technique for small samples while enhancing the interpretation of local DIF analyses. Preliminary analyses show that, the boundary correction implemented in KS-CATSIB, while useful in obtaining a less biased DIF estimate, it is not faring any better than CATSIB. Further examination of kernel smoothing adjustment for boundaries is imminent before it can be recommended.

Few corrections for improving the estiamtion in boundaries were considered. The first correction was not to estimate design points within a bandwidth distance of either boundary. This resulted in too many examinees not being used in the DIF estimation. To overcome this problem, the bandwidth was narrowed until it was small enough that at least 97.5% of either group examinees were included in the DIF estimation. The third modification made was to use the standard recommended bandwidth for as much of the ability range as possible, until the bandwidth bumped against the boundary. Then the bandwidth was allowed to shrink as necessary as it approached the boundary until at least 97.5% of either group was included in the DIF estimation. The last two modifications also did not produce satisfactory results. Habing (1999) evaluated Rice correction with a different boundary kernel due to Mueller (1991) that seems to be a promising. This is another method we will be evaluating in a future study. Subsequent to obtaining optimal smoothing corrections for boundaries, future studies will investigate Type I error rates and the degree to which kernel smoothing can increase the power of CATSIB for small sample sizes.

Because equity concerns have been especially of concern with computerized tests, it

is important that testing companies not allow their DIF assessment power to fall victim to the small sample sizes that CATs typically impose. First and foremost, the most powerful DIF detection possible should be implemented at the pretest stage. Secondly, the operational items should be continually monitored for DIF, using both the pretest and operational data in order to obtain maximum detection power. We hope that KS-CATSIB will be potentially a valuable tool in meeting these objectives.

References

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33,* 231-251.

Douglas, J., Stout, W., DiBello, L. (1996). A Kernel-smoothed version of SIBTEST with applications to local DIF inference and function estimation. *Journal of Educational and Behavioral Statistics, 21,* 333-363.

Habing, B. (1999, April). *Optimization of kernel smoothing for IRF estimation.* Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Mueller, H. G. (1991). Smooth otimum kernel estimators near endpoints. *Biometrika, 78,* 521-530.

Nandakumar, R., & Roussos, L.A. (in press). CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests. *LSAC Computerized Testing Report,* 97-11. Newtown, PA: Law School Admission Council.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Applications,* 9, 141-42.

Owen, R. J. (1969). *Bayesian approach to tailored testing.* (Research Bulletin 69-92). Princeton NJ: Educational Testing Service.

Ramsay, J., O. (1991). Kernel smoothing approach to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611-630.

Rice, J. (1984). Boundary modifications for kernel regression. *Commun. Statist.-Theor. Metho.,* 13(7), 893-900.

Roussos, L. A., Nandakumar, R., & Cwikla-Banks, J. (in press). Theoretical formula for statistical bias in CATSIB DIF estimator due to discretization of the ability scale. *LSAC Computerized Testing Report,* 1999. Newtown, PA: Law School Admission Council.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type 1 error performance. *Journal of Educational Measurement, 33*, 215-230.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Watson, G. (1964). Smooth regression analysis. *Sankhya A.*, 26, 359-372.

Table 1 – Item Parameter of DIF Items

Item	β	a	b_R	b_F
Equal Sizes for Reference and Focal Groups				
1	.00	.4	-1.50	-1.50
2	.05	.4	-1.74	-1.26
3	.10	.4	-1.98	-1.02
4	0	.4	1.50	1.50
5	.05	.4	1.26	1.74
6	.10	.4	1.02	1.98
7	.00	.8	0.00	0.00
8	.05	.8	-0.13	0.13
9	.10	.8	-0.25	0.25
10	.00	1.0	-1.50	-1.50
11	.05	1.0	-1.69	-1.31
12	.10	1.0	-1.88	-1.12
13	.00	1.4	1.50	1.50
14	.05	1.4	1.31	1.69
15	.10	1.4	1.12	1.99
16	.00	1.4	-1.50	-1.50
Reference Group Twice Size of Focal Group				
1	.00	.4	-1.5	-1.5
2	.05	.4	-1.66	-1.19
3	.10	.4	-1.81	-0.88
4	.00	.4	1.50	1.50
5	.05	.4	1.34	1.83
6	.10	.4	1.17	2.17
7	.00	.8	0.00	0.00
8	.05	.8	-0.08	0.17
9	.10	.8	-0.17	0.34
10	.00	1.0	-1.50	-1.50
11	.05	1.0	-1.62	-1.26
12	.10	1.0	-1.74	-1.03
13	.00	1.4	1.50	1.50
14	.05	1.4	1.37	1.77
15	.10	1.4	1.22	2.07
16	.00	1.4	-1.50	-1.50

Table 2: Comparison of average DIF estimates for KS-CATSIB With and Without Rice Correction (100 trials)

item #	True beta	KS-CATSIB (no Rice)	KS-CATSIB (Rice)
1	0.00	0.006 (.0392)	0.0043 (.0400)
2	0.05	0.0614 (.0486)	0.0571 (.0500)
3	0.10	0.1119 (.0446)	0.1066 (.0478)
4	0.00	0.0087 (.094)	0.0073 (.0521)
5	0.05	0.0544 (.0521)	0.0545 (.0543)
6	0.10	0.1057 (.0454)	0.1080 (.0492)
7	0.00	0.0076 (.0532)	0.0014 (.0530)
8	0.05	0.0515 (.0468)	0.0433 (.0469)
9	0.10	0.0979 (.0448)	0.0889 (.0459)
10	0.00	0.0138 (.0338)	.0094 (.0330)
11	0.05	0.0662 (.0351)	.0601 (.0341)
12	0.10	0.1231 (.0381)	.1131 (.0348)
13	0.00	0.0245 (.0449)	.0284 (.0494)
14	0.05	0.082 (.0392)	.0955 (.0438)
15	0.10	0.1301 (.0450)	.1184 (.0460)
16	0.00	0.0229 (.0274)	.0196 (.0276)

Table 3: Comparison of average DIF estimates for KS-CATSIB and CATSIB

		100 trials		400 trials	
item #	True beta	KS-CATSIB	CATSIB	KS-CATSIB	CATSIB
1	0.00	0.006 (.0392)	0.0029 (.0351)	0.0073 (.0411)	0.0026 (.0371)
2	0.05	0.0614 (.0486)	0.0546 (.0443)	0.0606 (.0446)	0.0543 (.0419)
3	0.10	0.1119 (.0446)	0.1048 (.0401)	0.1127 (.0429)	0.1069 (.0395)
4	0.00	0.0087 (.094)	0.002 (0.0468)	0.0116 (.0473)	0.0024 (.0457)
5	0.05	0.0544 (.0521)	0.0433 (.0514)	0.0575 (.0498)	0.0502 (.0475)
6	0.10	0.1057 (.0454)	0.097 (.0417)	0.1119 (.0482)	0.1051 (.0458)
7	0.00	0.0076 (.0532)	0.0011 (.0463)	0.0136 (.0486)	0.0056 (.0423)
8	0.05	0.0515 (.0468)	0.0481 (.0417)	0.0608 (.0466)	0.0572 (.0434)
9	0.10	0.0979 (.0448)	0.0989 (.0370)	0.1085 (.0480)	0.1092 (.0435)
10	0.00	0.0138 (.0338)	0.004 (.0268)	0.0166 (.0342)	0.0067 (.0275)
11	0.05	0.0662 (.0351)	0.0565 (.0262)	0.0682 (.0350)	0.0561 (.0275)
12	0.10	0.1231 (.0381)	0.1131 (.0295)	0.1236 (.0387)	0.1104 (.0298)
13	0.00	0.0245 (.0449)	0.0037 (.0404)	0.0261 (.0434)	0.0007 (.0396)
14	0.05	0.082 (.0392)	0.0459 (.0384)	0.0805 (.0435)	0.051 (.0412)
15	0.10	0.1301 (.0450)	0.0896 (.0376)	0.1318 (.0457)	0.0968 (.0409)
16	0.00	0.0229 (.0274)	0.0088 (.0216)	0.0213 (.0279)	0.0075 (.0215)